DOCUMENT RESUME

ED 337 041 FL 019 755

AUTHOR Navarrete, Cecilia; And Others
TITLE Informal Assessment in Educational Evaluation: Implications for Bilingual Education Programs.
INSTITUTION National Clearinghouse for Bilingual Education, Washington, DC.
SPONS AGENCY Office of Bilingual Education and Minority Languages Affairs (ED), Washington, DC.
PUB DATE 90
CONTRACT 289004001; T288003002
NOTE 28p.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Bilingual Education Programs; Comparative Analysis; Elementary Secondary Education; *Evaluation Methods; Formative Evaluation; *Informal Assessment; *Language Tests; *Second Language Instruction; Standardized Tests; *Student Evaluation
IDENTIFIERS Elementary Secondary Education Act Title VII

ABSTRACT

Given the controversy over the use of standardized tests that rely heavily on multiple-choice items reflecting the language, culture, and/or learning style of the middle class majority, arguments are advanced for the use of alternative, supplemental forms of assessment. Informal assessment is defined as techniques that can easily be incorporated into classroom routines and learning activities, and are identified as unstructured (e.g., writing samples, homework, journals, games, debates) or structured (e.g., checklists, close tests, rating scales, questionnaires, structured interviews). Guidelines for informal assessment are offered, including scoring procedures such as holistic or analytic procedures, general impression markings, or error patterns. Guidelines for using another method, student portfolios, are detailed. Guidance is also offered for the evaluation of programs funded under the Elementary and Secondary Education Act Title VII, including reporting assessment data. It is concluded that informal techniques are needed to provide the continuous, ongoing measurement of student growth needed for formative evaluation and for planning instructional strategies. Contains 23 references. (LB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED337041

NCBE

3

Informal Assessment in Educational Evaluation: Implications for Bilingual Education Programs

by

[redacted]

Cecilia Navarrete

Judith Wilde

Chris Nelson

Robert Martinez

Gary Hargett

National Clearinghouse for Bilingual Education

Program Information Guide Series

Summer 1990

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Gomez, Joel

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

BEST COPY AVAILABLE

FL019755

The National Clearinghouse for Bilingual Education (NCBE) is funded by the U.S. Department of Education's Office of Bilingual Education and Minority Languages Affairs (OBEMLA) and is operated under Contract No. 289004001 by The George Washington University's Center for the Study of Education and National Development, jointly with the Center for Applied Linguistics. The contents of this publication do not necessarily reflect the views or policies of the Department of Education, nor does the mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. Readers are free to duplicate and use these materials in keeping with accepted publication standards. NCBE requests that proper credit be given in the event of reproduction.

Introduction[1]

Central to the evaluation of any educational program are the instruments and procedures used to assess that program's effects. Many programs use commercially available standardized tests to measure academic achievement or language proficiency. There are good reasons for doing so. Standardized tests usually are administered annually by school districts, providing a ready source of achievement data. Test publishers provide information about the test's validity and reliability, fulfilling another requirement of evaluation. And, standardized test scores generally have been accepted by educators and the community.

However, recent research on student achievement has focused on problems associated with overreliance on standardized tests (e.g., Haney & Madaus 1989; Marston & Magnusson 1987; Pikulski 1990; Shepard 1989). Alternative approaches to assessing student progress have been suggested that address many of the problems associated with standardized tests (e.g., Marston & Magnusson 1987; Rogers 1989; Wiggins 1989; Wolf 1989). The purpose of this guide is to review some of the problems associated with standardized testing, describe alternative assessment approaches, and discuss how these approaches might be employed by bilingual educators to supplement the use of standardized tests.

Concerns with Standardized Testing

Criticisms of standardized tests seem to have grown in proportion to the frequency with which, and the purposes for which, they are used (Haney & Madaus 1989). Pikulski (1990) suggests that the greatest misuse of standardized tests may be their overuse. Many districts now administer such tests at every grade level, define success or failure of programs in terms of test scores, and even link teacher and administrator salaries and job security to student performance on standardized test performance. Three areas often criticized in regard to standardized tests are content, item format, and item bias. Standardized tests are designed to provide the best match possible to what is perceived to be the "typical" curriculum at a specific grade level. Because a *bilingual education* program is built on objectives unique to the needs of its students, many of the items on a standardized test may not measure the objectives or content of that program. Thus a standardized test may have low content validity for specific bilingual education programs. In such a situation, the test might not be sensitive to actual student progress. Consequently, the program, as measured by this test, would appear to be ineffective.

Standardized achievement tests generally rely heavily on multiple-choice items. This item format allows for greater content coverage as well as objective and efficient scoring. However, the response required by the format is recognition of the correct answer. This type of response does not necessarily match the type of responses students regularly make in the classroom, e.g., the production or synthesis of information. If students are not used to responding within the structure imposed by the item format, their test performance may suffer. On the other hand, students may recognize the correct form when it is presented as a discrete item in a test format, but fail to use that form correctly in communication contexts. In this case, a standardized test may make the student appear more proficient than performance would suggest.

Further, some tests have been criticized for including items that are biased against certain kinds of students (e.g., ethnic minorities, limited English proficient, rural, inner-city). The basis for this criticism is that the items reflect the language, culture, and/or learning style of the middle-class majority (Neill & Medina, 1989). Although test companies have attempted to write culture-free items, the removal of questions from a meaningful context has proved problematic for minority students.

Thus, there are strong arguments in favor of educators considering the use of alternative forms of assessment to supplement standardized test information. These alternate assessments should be timely, not time consuming, truly representative of the curriculum, and tangibly meaningful to the teacher and student. Techniques of informal assessment have the potential to meet these criteria as well as programmatic requirements for formative and summative evaluations. Validity and reliability are not exclusive properties of formal, norm-referenced tests. Informal techniques are valid if they measure the skills and knowledge imparted by the project; they are reliable if they measure consistently and accurately.

Defining Informal Assessment

"Formal" and "informal" are not technical psychometric terms; therefore, there are no uniformly accepted definitions. "Informal" is used here to indicate techniques that can easily be incorporated into classroom routines and learning activities. Informal assessment techniques can be used at anytime without interfering with instructional time. Their results are indicative of the student's performance on the skill or subject of interest. Unlike standardized tests, they are not intended to provide a comparison to a broader group beyond the students in the local project.

This is not to say that informal assessment is casual or lacking in rigor. Formal tests assume a single set of expectations for all students and come with prescribed criteria for scoring and interpretation. Informal assessment, on the other hand, requires a clear understanding of the levels of ability the students bring with them. Only then may assessment activities be selected that students can attempt reasonably. Informal assessment seeks to identify the strengths and needs of individual students without regard to grade or age norms.

Informal Assessment Techniques

Methods for informal assessment can be divided into two main types: unstructured (e.g., student work samples, journals) and structured (e.g., checklists, observations). The unstructured methods frequently are somewhat more difficult to score and evaluate, but they can provide a great deal of valuable information about the skills of the children, particularly in the areas of language proficiency. Structured methods can be reliable and valid techniques when time is spent creating the "scoring" procedures.

While informal assessment utilizes open-ended exercises reflecting student learning, teachers (and students) can infer "from the mere presence of concepts, as well as correct application, that the student possesses the intended outcomes" (Muir & Wells 1983, p. 95). Another important aspect of informal assessments is that they actively involve the students in the evaluation process—they are not just paper-and-pencil tests.

Unstructured Assessment Techniques

Unstructured techniques for assessing students can run the gamut from writing stories to playing games and include both written and oral activities. The range of possible activities is limited only by the creativity of the teacher and students. Table 1 on page 4 presents several illustrative unstructured assessments/techniques.

Structured Assessment Techniques

Structured assessments are planned by the teacher much more specifically than are unstructured assessments. As the examples listed and described in Table 2 on page 6 indicate, structured assessment measures are more varied than unstructured ones. Indeed, some of them are types of tests of one kind or another. In each case, definitely "right" and "wrong," "completed" or "not completed" determinations can be made. Consequently, the scoring of structured assessment activities is relatively easier compared to the scoring of unstructured assessment activities.

Table 1

Types of unstructured assessment techniques

Writing Samples

When students write anything on specific topics, their products can be scored by using one of the techniques described in Table 3. Other creative writing samples that can be used to assess student progress include newspapers, newsletters, collages, graffiti walls, scripts for a play, and language experience stories.

Homework

Any written work students do alone, either in class or in the home, can be gathered and used to assess student progress. With teacher guidance, students can participate in diagnosing and remediating their own errors. In addition, students' interests, abilities, and efforts can be monitored across time.

Logs or journals

An individual method of writing. Teachers can review on a daily, weekly, or quarterly basis to determine how students are perceiving their learning processes as well as shaping their ideas and strengths for more formal writing which occurs in other activities.

Games

Games can provide students with a challenging method for increasing their skills in various areas such as math, spelling, naming categories of objects/people, and so on.

Debates

Students' oral work can be evaluated informally in debates by assessing their oral presentation skills in terms of their ability to understand concepts and present them to others in an orderly fashion.

Brainstorming

This technique can be used successfully with all ages of children to determine what may already be known about a particular topic. Students often feel free to participate because there is no criticism or judgment.

Story retelling

This technique can be used in either oral or written formats. It provides information on a wide range of language-based abilities. Recall is part of retelling, but teachers can use it to determine whether children understood the point of the story and what problems children have in organizing the elements of the story into a coherent whole. This also can be used to share cultural heritage when children are asked to retell a story in class that is part of their family heritage.

Anecdotal

This method can be used by teachers to record behaviors and students' progress. These comments can include behavioral, emotional, and academic information. For instance, "Jaime sat for five minutes before beginning his assignment." These should be written carefully, avoiding judgmental words.

Naturalistic

Related to anecdotal records, this type of observation may take the form of notes written at the end of the day by a teacher. They may record what occurred on the playground, in the classroom, among students, or may just reflect the general classroom atmosphere.

Table 2

Types of structured informal assessments

Checklists

Checklists specify student behaviors or products expected during progression through the curriculum. The items on the checklist may be content area objectives. A checklist is considered to be a type of observational technique. Because observers check only the presence or absence of the behavior or product, checklists generally are reliable and relatively easy to use. Used over time, checklists can document students' rate and degree of accomplishment within the curriculum.

Cloze Tests

Cloze tests are composed of text from which words have been deleted randomly. Students fill in the blanks based on their comprehension of the context of the passage. The procedure is intended to provide a measure of reading comprehension.

Criterion-referenced Tests

Criterion-referenced tests are sometimes included as a type of informal assessment. This type of test is tied directly to instructional objectives, measures progress through the curriculum and can be used for specific instructional planning. In order for the test to reflect a particular curriculum, criterion-referenced tests often are developed locally by teachers or a school district. Student performance is evaluated relative to mastery of the objectives, with a minimum performance level being used to define mastery.

Rating Scales

This is an assessment technique often associated with observation of student work or behaviors. Rather than recording the "presence" or "absence" of a behavior or skill, the observer subjectively rates each item according to some dimension of interest. For example, students might be rated on how proficient they are on different elements of an oral presentation to the class. Each element may be rated on a 1 to 5 scale, with 5 representing the highest level of proficiency.

Questionnaires

A questionnaire is a self-report assessment device on which students can provide information about areas of interest to the teacher. Questionnaire items can be written in a variety of formats and may be forced-choice (response alternatives are provided) or open-ended (students answer questions in their own words). Questionnaires designed to provide alternative assessments of achievement or language proficiency may ask students to report how well they believe they are performing in a particular subject or to indicate areas in which they would like more help from the teacher. One type of questionnaire (which assumes that the student can read in the native language) requests that students check off in the first language the kinds of things they can do in English. For a questionnaire to provide accurate information, students must be able to read the items, have the information to respond to the items, and have the writing skills to respond.

Miscue Analysis

An informal assessment of strategies used by students when reading aloud or retelling a story. Typically, students read a grade-level passage (e.g., 250 words) while a judge follows along with a duplicate copy of the passage. The student may be tape recorded. Each time an error occurs, the judge circles the word or phrase. A description of the actual error can be taken from the tape after the session and analyzed for errors in pronunciation, sentence structure, vocabulary, use of syntax, etc. (see Goodman 1973).

Structured Interviews

Structured interviews are essentially oral interview questionnaires. Used as an alternative assessment of achievement or language proficiency, the interview could be conducted with a student or a group of students to obtain information of interest to a teacher. As with written questionnaires, interview questions could be forced-choice or open-ended. Because the information exchange is entirely oral, it is important to keep interview questions (including response alternatives for forced-choice items) as simple and to-the-point as possible.

Guidelines for Informal Assessment

In order to be effective, informal assessment activities must be carefully planned. With appropriate planning, they can be reliable and valid, and they can serve diagnostic purposes as well as formative and summative evaluation purposes within all types of bilingual education programs. General guidelines are presented here to ensure these qualities. These guidelines apply both to formal and informal assessments.

Validity and Reliability

Standardized tests often are selected because their technical manuals report validity and reliability characteristics. However, if the content of these tests does not match the instructional objectives of the project, their validity is negated. For example, many standardized tests include structural analysis skills as part of the reading or language arts sections. If a bilingual education project does not teach structural analysis skills, concentrating instead on the communicative aspects of reading/writing, such a test may not be valid for that particular project.

The validity of informal measures can be established by demonstrating that the information obtained from a given technique reflects the project's instructional goals and objectives. If, for example, the project is teaching communicative writing, a collection of holistically scored writing samples would be a valid measure. Therefore, a first step toward validating the use of informal assessment measures is a clear statement of curricular expectations in terms of goals and objectives.

Reliability, in its purest sense, refers to the ability of a measure to discriminate levels of competency among persons who take it. This is accomplished through the consistent application of scoring criteria. As with validity, the reliability of informal measures can be established by a clear statement of the expectations for student performance in the curriculum and ensuring that teachers apply consistent criteria based on those expectations. If the informal measures accurately represent students' progress, and if they accurately distinguish the differential progress made by individual students, they are reliable.

Scoring Procedures

Consideration has to be given to the reliability and validity of the scoring procedures used in assessment, both formal and informal. Among critical issues to be addressed are:

1. The validity of the judgment may be limited by the heavy dependency on the opinion of raters. To ensure high reliability, raters must be trained to meet a set criterion (e.g., when judging ten individuals, raters should rate eight of them similarly).

2. The scores must be specific to the learning situation. The scoring procedure must match the exercise or performance. To ensure this match, the purpose for assessment and the content to be assessed must first be decided. Agreement should also be sought on the descriptors developed for each scoring category to be used.

3. Scoring procedures may be time consuming. To ensure success, the commitment and support of project and school personnel must be sought. Training and practice must be offered to the raters.

Scoring procedures utilized in unstructured assessment activities can be used to:

- measure progress and achievement in most content areas;
- measure literacy skills such as oral, reading, and written production;
- develop summative and formative evaluations;
- make an initial diagnosis of a student's learning;
- guide and focus feedback on students' work;
- measure students' growth over time or for specific periods;
- determine the effectiveness of an instructional program;
- measure group differences between project students and nonproject comparison groups;
- analyze the performance of an individual student; and
- correlate student outcomes with formal, standardized tests of achievement and language proficiency.

Table 3 on page 10 lists some general scoring procedures and a brief summary description of popularly used techniques.

Combining Assessments for Evaluation

Different methods of combining types of structured and unstructured informal assessments and associated scoring procedures appear in the literature. While these approaches have different labels and differ somewhat in philosophy, all are offered as alternatives to standardized testing and use informal assessment to measure student performance in the context of the curriculum.

1. Curriculum-based assessment uses the "material to be learned as the basis for assessing the degree to which it has been learned"

Table 3

Scoring assessments for unstructured activities

Holistic

A guided procedure for evaluating performance (oral or written) as a whole rather than by its separate linguistic, rhetorical, or informational features. Evaluation is achieved through the use of a general scoring guide which lists detailed criteria for each score. Holistic judgments are made on the closest match between the criteria and the students' work. Criteria typically are based on a rating scale that ranges from 3 to 10 points (3 = low quality level and 10 = high quality level).

Primary Trait

A modified version of holistic scoring; the most difficult of all holistic scoring procedures, its primary purpose is to assess a particular feature(s) of a discourse or a performance (oral or written) rather than the students' work as a whole. Secondary level traits also can be identified and scored using this approach.

Analytic

A complex version of holistic scoring; students' work is evaluated according to multiple criteria which are weighted based on their level of importance in the learning situation. For example, a writing sample can be assessed on organization, sentence structure, usage, mechanics, and format. Each criterion is rated on a 1 to 5 scale (1 = low and 5 = high). A weighting scheme then is applied.

For example, the organization of an essay can be weighted six times as much as the format; sentence structure five times as much as format; and so on. This procedure can be used for many purposes such as diagnostic placement, reclassification and exiting, growth measurement, program evaluation, and educational research.

Holistic Survey

Uses multiple samples of students' written work representing three of five discourse modes: expressive, narrative, descriptive, expository, and argumentative. Prior to scoring, students select topics, repeat oral directions to demonstrate understanding of the task, and have the opportunity to revise and edit their work before submitting it for evaluation. The scoring procedures used in the survey can include primary trait, analytic, or other holistic scoring devices relevant to the goals and objectives of the written assignment.

General Impression Markings

The simplest of the holistic procedures. The raters score the papers by sorting papers along a continuum such as excellent to poor, or acceptable to unacceptable. Critical to this approach is that raters become "calibrated" to reach consensus by reading and judging a large sample of papers.

Error Patterns

The assessment of students' written work or mathematical computations. Scoring is based on a criterion that describes the process or continuum of learning procedures that reflect understanding of the skill or concept being assessed. A minimum of three problems or written assignments are collected and assessed to ensure that a student's error is not due to chance.

Assigning Grades

The "old standard." Students are assigned a number or letter grade based on achievement, competency, or mastery levels. Grades can be pass-fail or can reflect letter grades, such as A to F. The major limitation of this scoring procedure is that grades do not provide any information on the strengths or weaknesses in a content area.

(Tucker 1985, p. 199). This approach employs informal measures such as writing samples, reading samples from the basal series, and teacher-made spelling tests from the basal series. It has received a good deal of attention in the special education literature (e.g., Deno 1985; Marston & Magnusson 1987) and was developed, in part, in response to the need to address performance criteria specified in students' individualized education plans (IEPs).

2. Ecological assessment (e.g., Bulgren & Knackendoffel 1986) evaluates student performance in the context of the environment. Sources of such data include student records, student interviews, observations, and collections of student products. Ecological assessment takes into account such things as the physical arrangement of the classroom; patterns of classroom activity; interactions between the teacher and students and among students; student learning styles; and expectations of student performance by parents, peers, and teachers.

3. Performance assessment (Stiggins 1984) provides a structure for teachers to evaluate student behavior and/or products. Assessments can take any form, depending on the behavior or product of interest, and are designed according to four considerations: (1) a decision situation that defines the basic reason for conducting the assessment; (2) a test activity or exercise to which the student responds; (3) the student response; and (4) a rating or judgment of performance.

Student Portfolios

A method which can combine both informal and formal measures is portfolio assessment (e.g., Wolf 1989). This method is rapidly gaining in popularity because of its ability to assess student work samples over the course of a school year or even longer. For this reason a more detailed description of portfolios follows.

Portfolios provide an approach to organizing and summarizing student data for programs interested in student- and teacher-oriented assessments. They represent a philosophy that views assessment as an integral component of instruction and the process of learning. Using a wide variety of learning indicators gathered across multiple educational situations over a specified period of time, portfolios can provide an ecologically valid approach to assessing limited English proficient students. While the approach is not new, portfolios are useful in both formative and summative evaluations, which actively involve teachers and students in assessment.

Portfolios are files or folders containing a variety of information that documents a student's experiences and accomplishments. The type of information collected for a portfolio can consist of summary descriptions of accomplishments, official records, and diary or journal items. Summary descriptions of accomplishments can include samples of the student's writing; artwork or other types of creations by the students; and testimonies from others (e.g., teachers, students, tutors) about the student's work.

Formal records typically included in a portfolio are scores on standardized achievement and language proficiency tests; lists of memberships and participation in extracurricular clubs or events; lists of awards and recognitions; and letters of recommendation.

Diaries or journals can be incorporated in portfolios to help students reflect on their learning. Excerpts from a diary or journal are selected for the portfolio to illustrate the students' view of their academic and emotional development.

Valencia (1990) recommends organizing the content of the portfolios into two sections. In the first section, the actual work of the students, or "raw data," is included. The information in this section assists the teacher to examine students' ongoing work, give feedback on their progress, and provide supporting documentation in building an in-depth picture of the student's ability. The second section consists of summary sheets or organizational frameworks for synthesizing the student's work. The information summarized in the second section is used to help teachers look systematically across students, to make instructional decisions, and for reporting purposes.

One major concern in using portfolios is with summarizing information within and across classrooms in a consistent and reliable manner, an issue discussed below.

Guidelines for Using Portfolios in Bilingual Education Evaluations

As part of the bilingual education evaluation, the portfolios can be quite useful. They can:

- be used to meet many of the bilingual education evaluation requirements;
- involve both formal and informal assessment methods;
- offer a comprehensive view of students' academic achievement and linguistic proficiency;

- provide more detailed information on those aspects of students' performance which are not readily measured by traditional examining methods;
- reflect the taught curriculum and individual child's learning experiences;
- encourage teachers to use different ways to evaluate learning;
- document the student's learning and progress; and
- help teachers examine their own development and skills.

Although the shape and form of portfolios may change from program to program, the real value of a portfolio lies in three areas. In the first area, portfolios have the potential to provide project teachers and students with a rich source of information to understand the development and progress of project students and to plan educational programs that enhance student learning and "showcase" their achievements. In the second area, portfolios allow for reporting in a holistic and valid way. The information gathered in a portfolio is taken from actual student work and assessment focuses on the whole of what a student learns, not on discrete and isolated facts and figures. In the third area, formal and informal data can be used in a nonadversarial effort to evaluate student learning in a comprehensive and authentic manner.

Although portfolio assessment offers great flexibility and a holistic picture of students' development, several technical issues must be addressed to make portfolios valid for bilingual education evaluations. These issues are summarized in three organizational guidelines which are based on current research and instructional practices in education (Au, Scheu, Kawakami, & Herman 1990; Jongsma 1989; Pikulski 1989; Simmons 1990; Stiggins 1984; Valencia 1990; Wolf 1989).

1. Portfolios Must Have a Clear Purpose

To be useful, information gathered for portfolios must reflect the priorities of the program. It must be kept in mind that the purpose of a bilingual education program evaluation stems from the goals of the actual program. The first critical step, then, is to identify and prioritize the key program goals of curriculum and instruction. In developing goals for portfolio assessment, it will be helpful to review (a) the state's current language arts and bilingual curriculum guidelines, (b) the district's or state's standardized achievement and language proficiency tests, and (c) the scope and sequence

charts of the reading and literacy materials that will be used with the students.

Note that the goals of a program should be broad and general, not overly specific, concrete, or isolated lesson objectives. For example, a goal may be written as "To learn reading comprehension skills," or "To write fluently in English." If goals are too specific, portfolios can get cluttered with information that may not be useful to the student, teacher, administrator, or evaluator.

2. Portfolios Must Interact With the Curriculum

This issue also is known as content validity. It is important that the information in portfolios accurately and authentically represent the content and instruction of the program. Content validity can be maximized by making sure portfolios contain (a) a clear purpose of the assessment, (b) a close link between the behaviors or products collected and the evaluation goals, (c) a wide variety of classroom exercises or tasks measuring the same skill, and (d) a cross-check of student capabilities based on both formal tests and informal assessments.

When deciding on the type of assessment information to include in the portfolio, existing instructional activities should be used. Most likely, the information will be appropriate for portfolios. For example, one of the goals in the Kamehameha Elementary Education Program (KEEP) in Hawaii is to increase students' interest in reading and expand their repertoire of book reading. To determine to what extent this goal is achieved, teachers use a checklist to examine students' reading logs. The logs include a list of the titles and authors of the books students have read. With this information, teachers review each student's list in terms of level of appropriateness, genres read, and book preferences. Students also are asked to include dates the books were read in order to determine the number of books read over specified periods of time. The information thus obtained is then summarized in the checklist and used to monitor and report on students' learning as well as to improve instruction.

3. Portfolios Must Be Assessed Reliably

Reliability in portfolios may be defined as the level of consistency or stability of the devices used to assess student progress. At present, there are no set guidelines for establishing reliability for portfolios. The major reason is that portfolios, by their nature, are composed of a broad and varied collection of students' work from

oral reading, comprehension checks, and teachers' observation notes to formal tests of the students' achievement or proficiency. Equally important, large-scale portfolio assessment has only recently been investigated as an alternative device in educational evaluation and research (Brandt 1988; Burnham 1986; Elbow & Belanoff 1986; Simmons 1990; Wolf 1989).

However, there are several criteria which are recommended in estimating the reliability of portfolios for large-scale assessment. These criteria apply both at the classroom level and at the grade level. Teachers and administrators must, at a minimum, be able to

- design clear scoring criteria in order to maximize the raters' understanding of the categories to be evaluated;
- maintain objectivity in assessing student work by periodically checking the consistency of ratings given to students' work in the same area;
- ensure inter-rater reliability when more than one person is involved in the scoring process;
- make reliable and systematic observations, plan clear observation guidelines;
- use objective terminology when describing student behavior;
- allow time to test the observation instrument and its ability to pick up the information desired;
- check for inter-rater reliability as appropriate;
- keep consistent and continuous records of the students to measure their development and learning outcomes; and
- check judgments using multiple measures such as other tests and information sources.

A major issue that arises in the use of portfolios relates to the problem of summarizing data within and across classrooms in a consistent and reliable manner. Using the guidelines suggested above in the planning and organization of portfolios will provide for reliable and valid assessment. These guidelines, however, are only a framework for the assessment procedures and will need to be applied by teachers to determine their effectiveness and practicality.

Evaluation of ESEA Title VII–Funded Programs

Title VII of the Elementary and Secondary Education Act provides funding to school districts for implementing bilingual education programs to help limited English proficient students learn English. There is a requirement that each program receiving funding under Title VII submit yearly program evaluation results.

Title VII regulations focus on summative evaluation, the judgment of the effectiveness of a program. Formative evaluation, which provides feedback during a program so that the program may be improved, is also a concern. Informal assessment procedures can be used for both types of evaluation.

Informal measures are ideal for formative evaluation, because they can be given frequently and lend themselves to nearly immediate scoring and interpretation. To the extent that informal measures are embedded in the curriculum, they provide formative information as to whether the expected progress is being made. Where informal measures show that progress has been made, they confirm the decision to move students forward in the curriculum. Where they show that the expected progress has not been made, they suggest modification of the current approach or perhaps may call for a different instructional approach.

Informal measures also may be particularly appropriate for diagnostic assessment of individual students. As mentioned above, formal standardized tests may not necessarily focus on the skills that a specific group of limited English proficient students are being taught. Informal measures should be drawn directly from the work the class is engaged in and thus provide evidence of mastery of intended objectives. The teacher can examine each student's work for that evidence.

Informal measures tend to be production or performance measures. This means children are tested by actually doing whatever it is the teacher hopes they can do. For example, limited English proficient children often confuse she/he/it or leave off the "s" on third person singular English verbs (e.g., "she run" for "she runs"). An informal measure should demonstrate whether the child can produce the distinction between she/he/it or say "she runs." In contrast, most formal tests are indirect measures that ask the child to recognize a correct form (among several forms, some of which are "incorrect"). Recognition and production involve very different skills. Recognition of a linguistic distinction does not imply the ability to produce that distinction. Thus a formal measure might give an erroneous indication of a student's competence.

Informal assessments also can be used for summative evaluation reports. In general, three conditions allow for the use of informal assessment in summative evaluation. First, goals must be operationalized as clearly stated performances that can be measured. Second, informal measures must be selected and applied consistently and accurately in order to match the operationalized goals. And third, the measures must be scored in a way that permits the aggregation of individual scores into group data that represent performance vis-à-vis the stated goals. This means that either the assessments, the scoring procedures, or both must have uniformity across the students.

Reporting Assessment Data

Title VII evaluation regulations require that both formal and informal assessment data be summarized across students. These regulations allow for the collection of both qualitative and quantitative data. Descriptions of pedagogical materials, methods, and techniques utilized in the program certainly can be addressed using either qualitative or quantitative data. Reporting the academic achievement of project participants using valid and reliable measures essentially requires a quantitative approach.

Informal assessment of student achievement or language proficiency, when used to supplement standardized achievement test data, probably is approached best from a quantitative perspective. Quantitative data collected toward this end meets the current Title VII evaluation regulations for reporting student achievement and proficiency data and has the potential to be aggregated more readily across students. Efficiency is important in accumulating data for an evaluation. Data can be collected both for purposes of feedback to program personnel and for the evaluation reports submitted to the Office of Bilingual Education and Minority Languages Affairs (OBEMLA). Thus OBEMLA describes types of data to be collected, but formal versus informal assessment approaches are not prescribed. The data required can be summarized into three areas: student outcomes, program implementation, and technical standards. Program staff and evaluators should refer to the appropriate *Federal Regulations* for specific information.[2]

Student Outcome Data

In reporting the academic achievement and language proficiency outcomes of project students, formal and informal assessments can be combined to meet the federal evaluation regulations.

Information on formal assessment may indicate how well students are performing in relationship to other students across the nation, state, and/or school district as well as at the school and classroom level. In addition, reporting achievement scores by subscale (e.g., vocabulary, grammar, comprehension) rather than total scores (e.g, reading) provides a finer breakdown and understanding of students' strengths and weaknesses and pinpoints areas of improvement.

Synthesized informal data can be used to support formal test findings or to provide documentation of the students' progress in instructional areas not covered in a formal test. In addition, informal data can provide more specific information about student progress through the curriculum and can provide it continuously throughout the year. The key to using informal data is that the information pertains to program goals and related objectives. Informal data can answer questions such as: What skills or concepts did the student actually learn during the academic year? To what extent did students have the opportunity to acquire the particular skills or concepts? What progress did the students make over the year? How did the students' attitudes affect learning?

Formal or informal approaches can be used to address rates of change as long as the information on each participating student is maintained. The information also must be collected in a continuous and accurate manner.

An additional Title VII evaluation requirement is that project student outcomes be compared to those of a nonproject comparison group. In addressing this requirement, similar formal and informal assessment procedures should be utilized where possible. However, if access to a nonproject comparison group is limited, then information for project and nonproject groups should be provided at least on academic achievement, language proficiency, and, if available, rates of change in attendance, drop-out, and postsecondary enrollment. This data collection provides a valid comparison of the project students' learning outcomes and answers the question, "How do project students compare to similar students not receiving project support?"

Program Implementation

The essential purpose of evaluating the implementation of the program is to answer the question, "Does the unique combination of activities, instructional practices, materials, and role of the staff in the project lead to the achievement of its objectives?" Under Title

VII, information is required on program implementation including a description of instructional activities, time spent in those activities, and background on the staff responsible for carrying out the program.

One informal technique for collecting information is through existing information sources. In portfolios, for example, information can be collected on the students' backgrounds, needs, and competencies as well as on specific activities completed for children who may be handicapped or gifted and talented. Attendance lists also can be used in calculating the amount of time students received instructional services in the project. Information on the instructional time, specific educational activities, and instructional strategies can be collected and reported from teacher lesson plans or from teacher activity logs. The educational and professional data about the staff can be found in their job application forms. While this method can produce accurate information, a major concern in relying on this approach is that data collected may be incomplete or not relevant.

Another approach to use in collecting the required information is through the use of self-reports such as questionnaires and interviews. These methods of data collection can be used in two ways. First, information gathered may provide "recollected" or indirect versions of how the program was implemented. Since recollected data is relatively weak, also include supporting evidence, such as observations or existing records, whenever possible. On the other hand, these methods also can be used to collect information in an ongoing fashion which can result in more reliable data.

Technical Standards

For programs to have meaning they must have a standard point of reference. A standard is a set of baseline criteria that provides principles or rules for determining the quality or value of an evaluation. Title VII regulations require a description of specific technical standards in the annual evaluation report. These standards include a description of the data collection instruments and procedures, test administration and scoring, and accuracy of the evaluation procedures as well as the process for selecting a nonproject comparison group. When using either formal or informal assessment, describe how:

1. the nonproject group was selected;
2. conclusions apply to the persons, schools, or agencies served by the project;
3. instruments consistently and accurately measure progress toward accomplishing the objectives of the project; and

4. instruments appropriately consider factors such as age, grade, language, degree of language fluency and background of the persons being served.

The standards are intended to ensure that an evaluation conveys sound information about the features of the program. These standards require that the program information be technically adequate and that conclusions be linked logically to the data.

Conclusion

The pervasive theme in data collection is that bilingual education programs should strive to make their evaluations practical, viable, and accurate. By using a combination of both formal and informal assessments these requirements can be met effectively. We have not proposed that informal assessment be used in place of standardized tests; rather that they be used in conjunction with standardized tests.

While these formal measures provide general year-to-year progress of students in global content areas, they cannot provide the continuous, ongoing measurement of student growth needed for formative evaluation and for planning instructional strategies. Informal techniques can do so. The challenges faced in using informal assessment in the evaluation of bilingual education programs are the following:

- First, can informal assessment be held up to the same psychometric standards applied to formal assessment? With techniques such as those suggested above, reliable and valid informal assessment can be developed.

- Second, can further procedures be developed for aggregating the diverse information provided by informal assessment into a meaningful set of indices that allow us to state whether or not our programs are effective?

We believe these challenges can be met within bilingual education by using current understanding of informal assessment as a foundation on which to build.

Endnotes

1. This document has been produced by staff at the Evaluation Assistance Center (West) under contract #T288003002 with the U.S. Department of Education.

2. For those wishing to consult these regulations, see the Department of Education 34 CFR Part 500.50-500.52 as published in the *Federal Register* June 19, 1986 and October 5, 1988.

Bibliography

Au, K., Scheu, A., Kawakami, A., & Herman, P. (1990). Assessment and accountability in a whole literacy curriculum. *The Reading Teacher, 4,* 574–578.

Brandt, R. (1988). On assessment in the arts: A conversation with Howard Gardner. *Educational Leadership, 45* (4), 30–34.

Bulgren, J. A., & Knackendoffel, A. (1986). Ecological assessment: An overview. *The Pointer, 30* (2), 23–30.

Burnham, C. (1986). Portfolio evaluation: Room to breathe and grow. In Charles Bridges (Ed.), *Training the teacher of college composition.* Urbana, IL: National Council of Teachers of English.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52,* 219–232.

Elbow, P., & Belanoff, P. (1986). Portfolios as a substitute for proficiency examinations. *College Composition and Communication, 37,* 336–339.

Goodman, K. (1973). Analysis of oral reading miscues: Applied psycholinguistics. In F. Smith (Ed.), *Psycholinguistics and reading.* New York: Holt, Rinehart and Winston, Inc.

Haney, W., & Madaus, G. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. *Phi Delta Kappan, 70,* 683–687.

Jongsma, K. S. (1989). Portfolio assessment. *The Reading Teacher, 43,* 264–265.

Marston, D., & Magnusson, D. (1987). *Curriculum-based measurement: An introduction.* Minneapolis, MN: Minneapolis Public Schools.

Muir, S., & Wells, C. (1983). Informal evaluation. *Social Studies, 74* (3), 95–99.

Neill, D. M., & Medina, N. J. (1989). Standardized testing: Harmful to educational health. *Phi Delta Kappan, 70,* 688–697.

Pikulski, J. J. (1990). Assessment: The role of tests in a literary assessment program. *The Reading Teacher, 44,* 686–688.

Pikulski, J. J. (1989). The assessment of reading: A time for change? *The Reading Teacher, 43,* 80–81.

Rogers, V. (1989). Assessing the curriculum experienced by children. *Phi Delta Kappan, 70,* 714–717.

Shepard, L. A. (1989). Why we need better assessments. *Educational Leadership, 46* (7), 4–9.

Simmons, J. (1990). Portfolios as large scale assessment. *Language Arts, 67,* 262–267.

Stiggins, R. J. (1984). *Evaluating students by classroom observation: Watching students grow.* Washington, DC: National Education Association.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22,* 271–286.

Tucker, J. A. (1985). Curriculum-based assessment: An introduction. *Exceptional Children, 52,* 199–204.

Valencia, S. (1990). A portfolio approach to classroom reading assessment: The whys, whats, and hows. *The Reading Teacher, 43,* 338–340.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70,* 703–713.

Wolf, D. P. (1989). Portfolio assessment: Sampling student work. *Educational Leadership, 46*(7), 35–39.

About the Authors

The authors are on the staff of the Evaluation Assistance Center (West) at the University of New Mexico.

Cecilia Navarrete, Senior Research Associate, received her Ph.D. in Education from Stanford University.

Judith Wilde, Methodologist, received her Ph.D. in the Psychological Foundations of Education from the University of New Mexico.

Chris Nelson, Senior Research Associate, received her Ph.D. in Educational Psychology and Research from the University of Kansas.

Robert Martínez, Senior Research Associate, received his Ph.D. in Educational Research from the University of New Mexico.

Gary Hargett, Research Associate, is a doctoral candidate in Education at the University of Washington.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1990 273-553

The George Washington University
WASHINGTON DC

CAL
Center for Applied Linguistics